Online Appendix – NOT FOR PUBLICATION

"Fear, Appeasement, and the Effectiveness of Deterrence"

August 24, 2016.

This Online Appendix is divided into three parts. Appendix A contains proofs of the main propositions and corollaries text, as well as notes on Figures 3 and 4. Appendix B contains an expanded version of the case study on the Turkish Straits Crisis of 1946, additional notes on the Russo-Finnish War case study, and an additional case study on the Taiwan Straits Crisis of 1954-1955. Appendix C analyzes model variants and proves results discussed in the Robustness section.

A Proofs of Main Results

Proof of Proposition 1 The defender's strategy consists of a probability of responding to the transgression with war, which we denote α. The challenger's utility from not transgressing is n_C^1, and from transgressing is $\alpha \cdot w_C^1(\theta_C) + (1-\alpha) \cdot \max\left\{w_C^2(\theta_C), n_C^2\right\}$. The latter is strictly increasing in θ_C and greater than n_C^1 for all α when $\theta_C = \bar{\theta}_C^1$. Thus, the challenger's strategy must be to always transgress, or to transgress i.f.f her type is above a cutpoint $\bar{\theta}_C \leq \bar{\theta}_C^1$ at which she is indifferent between transgressing and not.

The necessary and sufficient conditions for existence of the two pure strategy equilibria ($\alpha^* = 0$ the no deterrence equilibrium, and $\alpha^* = 1$ the deterrence equilibrium) are described in the main text and straightforward to derive. There may also exist mixed strategy equilibria in which the defender responds with war with a strictly interior probability $\alpha^* \in (0,1)$. For such an equilibrium to hold, the defender must be indifferent between responding with war and allowing the transgression. This requires that,

$$P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C\right) = \bar{\beta} \tag{1}$$

i.e. the defender's posterior belief that the challenger will initiate war if allowed to transgress is equal to his threshold belief $\bar{\beta}$. The left hand side approaches $P\left(\theta_C \geq \bar{\theta}_C^2\right)$ as $\bar{\theta}_C$ approaches the lower bound of the type space, is equal to 1 at $\bar{\theta}_C = \bar{\theta}_C^2$, and is strictly increasing in between. Thus, a cutpoint satisfying (1) exists i.f.f. the no deterrence equilibrium exists ($P\left(\theta_C \geq \bar{\theta}_C^2\right) < \bar{\beta}$). We denote this cutpoint $\bar{\theta}_C^*$, which must be $< \bar{\theta}_C^2$.

We now check conditions such that there exists some $\alpha^* \in (0,1)$ that induces the challenger to play the cutpoint strategy $\bar{\theta}_C^* < \bar{\theta}_C^2$. A necessary condition and sufficient condition is that this type be indifferent between transgressing and not, i.e. there exists an α^* s.t.

$$\alpha^* \cdot w_C^1\left(\bar{\theta}_C^*\right) + (1-\alpha^*) \cdot n_C^2 = n_C^1. \tag{2}$$

If $\bar{\theta}_C^* > \bar{\theta}_C^1 \iff P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C^1\right) < \bar{\beta}$ (i.e. the deterrence equilibrium does not exist) then the condition cannot be satisfied since this would imply that both $w_C^1\left(\bar{\theta}_C^*\right)$ and n_C^2 are greater than n_C^1. Conversely, if $\bar{\theta}_C^* < \bar{\theta}_C^1$ then an α^* satisfying (2) exists and is unique.

Thus, a unique mixed strategy equilibrium exists i.f.f. both the no deterrence and deterrence equilibria exist, and the equilibrium strategies $\left(\alpha^*, \bar{\theta}_C^*\right)$ are uniquely characterized by (1) and (2). We now show that when there are multiple equilibria, i.e. $\bar{\beta} \in \left[P\left(\theta_C \geq \bar{\theta}_C^2\right), P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C^1\right)\right]$, the defender is strictly better off in the deterrence equilibrium than in either the no deterrence or mixed strategy equilibrium. The defender's utility in the deterrence equilibrium is $U^{de} = P\left(\theta_C < \bar{\theta}_C^1\right) \cdot n_D^1 + P\left(\theta_C \geq \bar{\theta}_C^1\right) \cdot w_D^1$. His utility in the mixed strategy equilibrium is

$$\begin{aligned} U^{ms} &= P\left(\theta_C < \bar{\theta}_C^*\right) \cdot n_D^1 + P\left(\theta_C \geq \bar{\theta}_C^*\right) \cdot \left(P\left(\theta_C < \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C^*\right) \cdot n_D^2 + P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C^*\right) \cdot w_D^2\right) \\ &= P\left(\theta_C < \bar{\theta}_C^*\right) \cdot n_D^1 + P\left(\theta_C \geq \bar{\theta}_C^*\right) \cdot w_D^1 \qquad \text{by def'n of } \bar{\theta}_C^*. \end{aligned}$$

This is less than U^{de} since $\bar{\theta}_C^* < \bar{\theta}_C^1$ by construction $\rightarrow P\left(\theta_C < \bar{\theta}_C^*\right) < P\left(\theta_C < \bar{\theta}_C^1\right)$, and $n_D^1 > w_D^1$. Finally, his utility in the no deterrence equilibrium is

$$\begin{aligned} U^{nd} &= P\left(\theta_C < \bar{\theta}_C^2\right) \cdot n_D^2 + P\left(\theta_C \geq \bar{\theta}_C^2\right) \cdot w_D^2. \\ &= P\left(\theta_C < \bar{\theta}_C^1\right) \cdot \underbrace{\left(P\left(\theta_C < \bar{\theta}_C^2 \,|\, \theta_C < \bar{\theta}_C^1\right) \cdot n_D^2 + P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C < \bar{\theta}_C^1\right) \cdot w_D^2\right)}_{< n_D^1 \text{ since } n_D^1 > n_D^2 > w_D^1 > w_D^2} \\ &\quad + P\left(\theta_C \geq \bar{\theta}_C^1\right) \cdot \underbrace{\left(P\left(\theta_C < \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C^1\right) \cdot n_D^2 + P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C^1\right) \cdot w_D^2\right)}_{< w_D^1 \text{ since the deterrence equilibrium exists}} \\ &< P\left(\theta_C < \bar{\theta}_C^1\right) \cdot n_D^1 \; + P\left(\theta_C \geq \bar{\theta}_C^1\right) \cdot w_D^1 = U^{de} \quad \blacksquare \end{aligned}$$

Proof of Proposition 2 If the defender knew the challengers type, then he would respond with war i.f.f. $\theta_C > \bar{\theta}_C^2$, and thus the challenger would be deterred i.f.f. $\theta_C \in \left(\bar{\theta}_C^2, \bar{\theta}_C^1\right)$. The probability of deterrence would therefore be $P\left(\theta_1 < \bar{\theta}_C^1, \theta_1 \geq \bar{\theta}_C^2\right)$. Now suppose first that the deterrence equilib-

rium exists when the challenger's type is unknown; then the probability of deterrence is $P\left(\theta_1 < \bar{\theta}_C^1\right)$, which is $>$ the probability of deterrence $P\left(\theta_1 < \bar{\theta}_C^1, \theta_1 \geq \bar{\theta}_C^2\right)$ when the challenger's type is known. Next suppose that the deterrence equilibrium does not exist when the challenger's type is unknown, so that the probability of deterrence is 0. Then we must have $\bar{\theta}_C^2 > \bar{\theta}_C^1$, which implies that the probability of deterrence is also 0 when the challenger's type is known.

Now we consider when the defender is better off not knowing the challenger's type. This is clearly the case when $\bar{\theta}_C^2 \leq \bar{\theta}_C^1$; types $< \bar{\theta}_C^2$ are deterred when they otherwise would not be, and for all other types the outcome is identical. So suppose that $\bar{\theta}_C^1 < \bar{\theta}_C^2$, and observe that a *necessary* condition for the defender to be better off not knowing is that the deterrence equilibrium exists. Hence we must characterize when the defender prefers not knowing and the deterrence equilibrium to knowing the challenger's type; this will be the case when

$$P\left(\theta_C < \bar{\theta}_C^1\right) \cdot \left(n_D^1 - n_D^2\right) > P\left(\theta_C \in \left[\bar{\theta}_C^1, \bar{\theta}_C^2\right)\right) \cdot \left(n_D^2 - w_D^1\right),$$

i.e. when the benefit $n_D^1 - n_D^2$ of deterring types $< \bar{\theta}_C^1$ exceeds the cost of preventable wars $n_C^2 - w_C^1$ against appeasable types $\theta_C \in \left[\bar{\theta}_C^1, \bar{\theta}_C^2\right)$. It is simple to show that the conjunction of this condition and the deterrence equilibrium existence condition reduces to the condition in the Proposition. ■

Proof of Corollary 2 For the purposes of expositional simplicity, we consider the game form in which the defender *first* chooses whether he will be informed or ignorant about the challenger's type, and nature *next* draws that type. (This is strategically equivalent to the game form in which nature moves first – since nature is non-strategic – but allows us to simply refer to proper subgames in which the defender is ignorant vs. informed). The result then follows immediately; by Proposition 1 the best equilibrium for the defender in the case of multiplicity involves selecting the deterrence equilibrium in the "ignorant" subgame whenever it exists, and the first stage is then simply the defender choosing which game he wants to play according to the calculus in Proposition 2. ■

Proof of Corollary 3

$$\begin{aligned} \delta_C^m\left(\bar{\theta}_C^1\right) \quad &\geq \quad \delta_C^d \iff w_C^1\left(\bar{\theta}_C^1\right) \geq n_C^1 + \left(\delta_C^d - \delta_C^m\left(\bar{\theta}_C^1\right)\right) \iff w_C^2\left(\bar{\theta}_C^1\right) \geq n_C^2 \\ &\iff \quad \bar{\theta}_C^2 \leq \bar{\theta}_C^1 \iff P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \bar{\theta}_C^1\right) = 1 \quad \blacksquare \end{aligned}$$

Proof of Proposition 3 Holding $\bar{\theta}_C^1$ (i.e. the challenger's first period payoffs) fixed, the ineffectiveness of appeasement $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right)$ is decreasing in $\bar{\theta}_C^2$. Thus for the first claim, it suffices to show $\bar{\theta}_C^2$ is decreasing in $\delta_C^m - \delta_C^d$. By assumption, $n_C^2 = n_C^1 + \delta_C^d$ and $w_C^2\left(\theta_C\right) = w_C^1\left(\theta_C\right) + \delta_C^m$ and $n_C^2 = w_C^2\left(\bar{\theta}_C^2\right)$, which together imply that $n_C^1 = w_C^1\left(\bar{\theta}_C^2\right) + \left(\delta_C^m - \delta_C^d\right)$. This implies the desired property since $w_C^1\left(\theta_C\right)$ is increasing in θ_C.

To show that the probability of deterrence is increasing in $\delta_C^m - \delta_C^d$, note that with the assumed equilibrium selection and by Proposition 1, the probability of deterrence is 0 if $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right) < \bar{\beta}$ and $P\left(\theta_C < \bar{\theta}_C^1\right)$ otherwise. Holding $\bar{\beta}$ (the defender's payoffs) fixed, the probability of deterrence is therefore (step-wise) increasing in $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right)$. Since this is increasing in $\delta_C^m - \delta_C^d$ the result is shown. $\blacksquare$

Construction of Figures 3 and 4 In Figure 3, the right panel depicts $G\left(\frac{1-F\left(\bar{\theta}_C^2\right)}{1-F\left(\bar{\theta}_C^1\right)}\right) \cdot F\left(\bar{\theta}_C^1\right)$, where $G\left(\bar{\beta}\right)$ is the induced probability distribution over $\bar{\beta}$. To generate both figures we assume that $w_C^1\left(\theta_C\right) = \theta_C$, the transgression's military and direct cost to the defender's are equal to .1, and the challenger's type is uniformly distributed on $[0, 1]$. The left panel assumes that the defender's benefit from peace is $n_D^1 - w_D^1 = .55$, while the right panel assumes it is uniformly distributed on $[-.1, 1]$.

Figure 4 is generated by assuming that $w_C^1\left(\theta_C\right) = \theta_C$ with $\theta_C \sim U[0, 1.1]$, and both the defender's benefit $n_D^1 - w_D^1$ and lowest type of challenger's benefit $n_C^1 - w_C^1\left(0\right)$ for avoiding war is .5. $\blacksquare$

B Case Study Notes

Extended Case Study: The Turkish Straits Crisis of 1946

We examine in depth a particular instance of deterrence success upon which the model sheds light; the crisis over control of the Turkish Straits that occurred between the United States and the Soviet Union in the early days of the Cold War. In 1945 and 1946, the Soviet Union repeatedly demanded that Turkey allow it to place bases on the Turkish Straits (Kuniholm 1980). These demands, coupled with extensive Soviet military preparations in the Balkans, led American officials to prepare for armed aggression against Turkey. In August 1946, President Truman decided that the United States would fight a full-scale war to defend Turkey in the event of Soviet invasion. Although this commitment was never announced publicly, it was communicated to Stalin through multiple channels. Upon learning about Truman's decision, Stalin reversed course (Mark 2005, pp. 123-124). We argue that the model we present helps explain why the United States was willing to fight, and why the Soviet Union found this threat credible.

Incentives The Turkish Straits Crisis contained the elements that are essential to our model: the defender's unwillingness to fight for the immediate stakes, the fear and uncertainty about the challenger's intentions, and the defender's preference to fight sooner rather than later. The United States had limited economic and political ties to Turkey and attached little intrinsic value to the Turkish Straits or Turkish independence.[17] The Soviet Union, on the other hand, had a direct interest in controlling the Straits. It sought military control of the Straits for the purposes of protecting trade and denying their use by hostile powers, objectives that American officials sympathized with.[18]

Furthermore, decision makers anticipated that any war with the Soviet Union would be enor-

[17]The U.S. had no obvious economic or political interests other than a small trade in tobacco, machinery and vehicles (Kuniholm 1980, pp. 65-66). The administration later claimed an interest in democratization, but this was merely rhetoric to justify the alliance to the public (Kayaoglu 2009).

[18]See DeLuca 1977 pp. 511-514; *FRUS 1946 VII*, 827-829; and Leffler (1985) pp. 809-810.

mously costly, despite the U.S. monopoly in atomic weapons. The military anticipated that the Red Army would launch offensives in Europe, the Middle East and Asia, that bombing would to be slow to produce results, and that ground operations would eventually be necessary (Ross 1996, pp. 12-19, 31). This combination of low stakes and a costly war makes the United States' decision to fight for Turkey puzzling. In fact, before the U.S. began to fear a general war with the Soviets, it made no plans for Turkey's defense despite believing that an attack was likely (Mark 1997, 398)

Fear By 1946, American officials had come to believe that the Soviet Union desired to dominate the Eurasian continent and eventually the world (Ross 1996, p. 3, 7). These ambitions did not necessarily imply that war would occur: most intelligence and military assessments assumed the Soviet Union was practical enough to avoid a destructive war with the United States and would accept the status quo (Mark 1997, 397). However, officials could not be perfectly confident in this assessment, and entertained the possibility that the Soviets would initiate general war in pursuit of their objectives. For example, in a meeting on June 12, 1946, President Truman speculated that the Soviet Union might start a war to divert public unrest, Secretary of the Navy Forrestal argued they might start a war if external circumstances were favorable for completing the "world revolution," and Admiral Leahy responded that the Soviets were simply unpredictable (Mark 2005, p. 119, 129).

If such a war were to occur, it is clear that the United States would have preferred that the fighting begin before losing Turkey. In the war plans, Turkey was to be the first line of defense against a Soviet advance toward strategically vital areas of the Middle East, the loss of which would weaken the U.S. and its allies. Turkish resistance to a Soviet offensive would protect American access to the Suez Canal, the Persian Gulf, and air bases in Egypt from which the United States planned to bomb central Russia (Leffler 1985, pp. 814-815).

Some officials believed that a successful transgression would not only strengthen the Soviet Union, but increase the Soviet appetite for general war. For example, Ambassador Edwin Wilson argued that, if the Soviet Union were allowed to overrun Turkey, they would be unable to resist the temp-

tation to advance toward the Suez Canal and the Persian Gulf. He wrote that "once this occurs, another world conflict becomes inevitable" because of the military advantages the Soviets would then have against the West (*FRUS 1946 VII*, p. 819, 822). This is similar to the unappeasability condition in the model: conquering Turkey would increase Soviet military strength and make general war more attractive, outweighing any pacifying effect from satisfying the Soviet demands over Turkey itself. Therefore, a concession could not appease the Soviet government if it was already belligerent.

Inference and Deterrence The Turkish Straits Crisis thus contained the incentive structure and uncertainty about challenger intentions that are essential to our model. Historical accounts of the crisis also suggest that, following the U.S. decision to defend Turkey, the Truman Administration was prepared to infer far-reaching Soviet ambitions from their willingness to attack Turkey *in the face of a U.S. commitment*, which is the key inference that sustains deterrence in equilibrium.

Most officials believed that the Soviet Union would be deterred by a U.S. commitment because it was generally thought that they wanted to avoid a major war (Mark 1997, 399). Conversely, officials appear to have believed that the Soviet Union would only invade Turkey if it did desire such a war. Undersecretary of State Dean Acheson said he believed that the Soviet Union would most likely be deterred, but he also argued the United States would "learn whether the Soviet policy includes an *affirmative* provision to go to war *now*" if deterrence failed.[19] This is the key inference in the deterrence equilibrium; the defender can learn of the challenger's preference for war from her willingness to transgress in the face of a deterrent threat. It is also clear that this inference sustained the U.S. willingness to initiate war following a Soviet attack. President Truman, when asked if he understood the decision to defend Turkey may mean war, responded that "we might as well find out whether the Russians were bent on world conquest now as in five or ten years" (Mills 1951, p. 192). Truman was prepared to infer far-reaching, long-term ambitions for world conquest from a Soviet willingness to challenge a credible U.S. commitment to Turkey.

[19]See Mark 1997, p. 400. Emphasis in original.

It is unclear whether the Soviets were deterred because they understood that American decision makers would interpret invasion as evidence of an intent to initiate war. As the first postwar crisis where the Soviets attempted to control an area where they did not already have a presence at the end of WWII, it seems likely that Stalin would have realized that invading Turkey would appear to the Americans as a dangerous new direction in Soviet policy, and that both parties ultimately came to understand the act of invasion as focal for revealing Soviet intentions. Although the invasion didn't occur, this episode dramatically reshaped American perceptions of Soviet intentions, and Soviet Foreign Minister V.M. Molotov later admitted that they had overreached (Mark 1997, 414).

Alternative Explanations While other deterrence mechanisms may have also been relevant, some fail to explain key features of the crisis. It is possible that reputational concerns drove decision-making, and that the United States felt it had to demonstrate its resolve to its allies and the Soviet Union. However, the primary fear in losing Turkey was never reputational, it was strategic. Government officials and military estimates repeatedly emphasized that the major concern in the crisis was that losing Turkey would disadvantage the United States in a future war with the Soviet Union. Truman himself, when told that a commitment to Turkey may mean war, pulled out a map and lectured his advisors about the strategic importance of the Middle East (Acheson 1969, 196).

Other commonly cited mechanisms in the deterrence literature do not seem to explain the outcome of the crisis. Any explanation involving audience costs would require threats to have been made publicly. While the United States did dispatch a naval force to the Mediterranean, it never publicly announced its decision to defend Turkey, instead communicating with the Soviet Union privately and downplaying the crisis in public (Trachtenberg 2012, 24-25). In addition, there was no obvious commitment device that would have automatically engaged the United States in a conflict, such as military forces stationed in Turkey as a "trip-wire." Finally, there was nothing probabilistic about the Americans' threat that "left something to chance." On the contrary, Truman clearly asserted that, if the Soviet Union invaded, he would follow the recommendation to defend Turkey "to the

end" (*FRUS 1946 VII*, 840).

Notes on Russo-Finnish War Case Study, 1939

In the literature on this case, the puzzle is not why deterrence failed, but why Finland was willing to fight rather than concede. Of particular interest is how the model here contributes to this existing debate. Consistent with the model, a common interpretation of Finland's decision to fight is that Finland feared that the granting of bases to the Soviets would weaken them in a future war (Jakobson 1961, 138-139; Van Evera 1999, 188). Sechser (2010) objects to this interpretation, arguing that these concessions were of limited military value and therefore can't explain Finland's willingness to fight against its far more powerful neighbor. Our model provides an explanation. It is clear that these military bases had more military value than intrinsic value, which would ensure they could not appease an already belligerent Soviet Union. While the case is not often examined from the perspective of deterrence, the information about Soviet intentions that Finland gained from the deterrence failure explains why they decided to fight rather than appease, thus demonstrating an important link between the circumstances of a deterrence failure and the outbreak of war.

Extra Case Study of Deterrence Success: Taiwan Straits Crisis, 1954-55

In 1954-55, the United States successfully deterred a Chinese invasion of the tiny island of Quemoy. Control of the island had some minor intrinsic value for the Communists: while it was a small island of mostly farmers and fisherman, Communist control would stop the Nationalists' occasional harassment of the mainland and merchant and fishing fleets near Amoy (Soman 2000, 120-121; Chang 1988, 99). However, its primary value was military: it housed a large contingent of Nationalist forces, blocked military deployments from Amoy toward Taiwan, and contained radar installations that helped defend Taiwan (Soman 2000, 120-125; Zhang 1992, 207-208). George and Smoke (1974) argue that control of the island had some "prestige" value. However, Communist propaganda during

this period focused mostly on Taiwan, and the attacks on Quemoy were launched soon after the beginning of a massive propaganda campaign about the liberation of Taiwan from the Nationalists. Quemoy's value in relation to the defense of Taiwan – as well as the importance of the U.S. fear of a Chinese invasion of Taiwan – can be seen in the fact that both Dulles and Eisenhower voiced a willingness to abandon Quemoy were China to pledge not to invade Taiwan (Wang 2011, 155; *FRUS 1955-1957 II*, pp. 146, 439). Nevertheless, the military importance of the island was not great, and certainly was not worth fighting a war over on its own. The island was distant from Taiwan and difficult to defend, and a Chinese invasion of Taiwan would have been extremely difficult with or without Quemoy due to American naval superiority in the Taiwan Straits. As a result, there were disagreements about its value within the Eisenhower Administration and between the U.S. and British governments (Zhai 1994, 159-161; Chang 1988, 100; Wang 2011, 172-173). The relatively small value of the island is also apparent in Eisenhower's worries about the difficulty of explaining to the American people why it would be worth starting a war with China over this small outpost (Zhai 1994, 159).

The model suggests that the deterrence of an Chinese invasion was successful because the island had *greater* military value than direct value, and therefore could not possibly appease a China intent on war. Given U.S. fears that China intended aggression against Taiwan, an invasion of Quemoy could have easily been perceived as an informative signal of both the present and future belligerence of the Communists (Zhang 1992, 193-194). The logic of the model, that attacking Quemoy would have signaled an affirmative desire for war *because* U.S. threats were credible, seemed to be operating. Communist propaganda acknowledged that the United States would defend Quemoy if attacked, and this was noted by American officials (Wang 2011, 180). In the face of a credible U.S. threat, the Communists backed down because of their desire to avoid war with the United States (Sheng, 487). Notably, the effectiveness of the U.S. threat over Quemoy is in contrast to the U.S. unwillingness to defend other island groups in the region like the Tachens, which had no military value for protecting

Taiwan, and which were promptly evacuated when attacked by Communist forces (Chang 1988, 102).

Extra Bibliography

Acheson, Dean. 1969. *Present at the Creation.* New York: Norton.

Chang, Gordon H. 1988. "To the Nuclear Brink: Eisenhower, Dulles and the Quemoy-Matsu Crisis." *International Security* 12(4):96-122.

DeLuca, Anthony R. 1977. "Soviet-American Politics and the Turkish Straits." *Political Science Quarterly* 92(3): 503-524.

George, Alexander L. and Richard Smoke. 1974. *Deterrence in American Foreign Policy: Theory and Practice.* New York: Columbia University Press.

Mark, Eduard. 1997. "The War Scare of 1946 and its Consequences." *Diplomatic History* 21(3):383-415.

Sheng, Michael M. 2008. "Mao and China's Relations with the Superpowers in the 1950s." *Modern China* 34(4):477-507.

Trachtenberg, Marc. 2012. "Audience Costs: An Historical Analysis." Security Studies 21(1):3-42.

U.S. Department of State. 1986. "Foreign Relations of the United States, 1955-57. Vol. II, China." United States Government Printing Office, Washington, DC.

Van Evera, Stephen. 1999. *Causes of War.* Ithaca, New York: Cornell University Press.

Wang, Tao. 2011. Isolating the Enemy: US-PRC Relations, 1953-1956. PhD Thesis, Georgetown University.

Zhai, Qiang. 1994. *The Dragon, the Lion and the Eagle: Chinese-British-American Relations, 1949-1958.* Kent, OH: Kent State University Press.

Zhang, Shu G. 1992. *Deterrence and Strategic Culture.* Ithaca, NY: Cornell University Press.

C Analysis of Model Variants in Robustness Section

C.1 Robustness to Salami Tactics and Endogenous Demands

In this section we consider robustness to two alternative bargaining protocols – a) extending the sequence so that the game resembles a model of salami tactics, and b) endogenizing the demand made by the challenger. In the former extension the defender always has the final move in each period over whether to fight or concede.

Rather than fully solve out general versions of these games, we present two examples illustrating that our basic insight holds in these variants. Both examples are constructed from the following payoff environment for a finite period game of conflict over a landmass of size and value equal to 1. In both variants there is no discounting and no "flow" payoffs – payoffs are based on the holdings of the landmass in the period in which the game ends.

Payoff Environment Suppose a challenger and a defender jointly occupy a landmass of size and value equal to 1. Say the **advantaged** party at time t is that which holds a majority of the landmass, and let δ_t denote the **excess** holdings of the advantaged party in period t above $\frac{1}{2}$. If a war occurs in period t, the probability the advantaged party wins is:

$$p\left(\delta_t\right) = \left(\frac{1}{2} + \delta_t\right) + \phi\left(\delta_t\right)$$

where $\phi\left(\delta_t\right) = \frac{2\delta_t(1-2\delta_t)}{Z}$ and Z is very large.[20] Also suppose that the defender's cost of war is commonly known to be $c_D \geq \frac{1}{4}$. The challenger's type θ_C is unknown and uniformly distributed over $\theta_C \sim U\left[-\frac{1}{4}, 0\right]$, and her cost of war is $c_C = -\theta_C$. ■

The challenger's probability of victory in a war as a function of her position is depicted in Figure 5 in the main text. We now present the first extension.

[20] We require at least $Z > 6$ for $p\left(\delta_t\right)$ to be strictly increasing in δ_t.

Extension 1 (Salami Tactics)**.** *Consider a $T \geq 3$ period game, and a $T+1$-length series of increasing values $0 < \delta_1 ... < \delta_T = \frac{1}{2}$. Each δ_t represents the challenger's* ***excess*** *holdings above $\frac{1}{2}$ if the game advances to period t. Assume that the challenger is initially advantaged ($\delta_1 > 0$), and that the increments of advancement $\delta_t - \delta_{t-1}$ are less than the defender's cost of war c_D for all $t < T$. In each period t, the challenger decides whether or not to attempt to advance from δ_t to δ_{t+1}. If she doesn't attempt to advance the game ends. If she does attempt to advance, the defender chooses whether or not to respond with war, and the challenger's probability of victory is $p(\delta_t)$.*

In this extension there are a finite number of exogenously fixed positions to which the challenger can advance, each advancement represents a transgression of fixed size, and each increment $\delta_t - \delta_{t-1}$ of advancement short of possessing the entire landmass is less than the defender's cost $c_D \geq \frac{1}{4}$ of war. Thus, the defender is always vulnerable to "salami tactics." When $\delta_1 < \delta_2 < c_D < \delta_3$, the game essentially reduces to the baseline model; the reason is that the defender knows she will be unable to credibly resist any advancement beyond δ_2, anticipates that conceding at δ_2 will result in a concession of size $1 - \delta_2 > c_D$, and is therefore there willing to fight.[21]

The set of equilibria for this extension satisfies the following proposition.

Proposition C.1. *If $c_D \in \left[\frac{1}{4}, \frac{1}{2}\right)$, then for any t^* such that $c_D < \frac{1}{2} - \delta_{t^*} \iff \delta_{t^*} < \frac{1}{2} - c_D$, there exists an equilibrium in which all types of challengers advance to δ_{t^*}, only challengers with cost $c_C < \phi(\delta_{t^*})$ attempt to advance to δ_{t^*+1}, and the defender always responds with war. If $c_D > \frac{1}{2}$ then in any equilibrium the challenger occupies the entire landmass.*

Proof: We first show the desired equilibrium when $c_D \in \left[\frac{1}{4}, \frac{1}{2}\right)$ and $\delta_t^* < \frac{1}{2} - c_D$. Define $\bar{t}$ as the period with the largest $\delta_{\bar{t}}$ strictly less than $\frac{1}{2} - c_D$, and observe that $\delta_{\bar{t}} < \frac{1}{4}$. Now consider the following strategy profile. After all histories, in periods $t \in [t^*, \bar{t}]$ the defender responds to

[21] When $\delta_1 < \delta_2 < c_D < \delta_3$, the game maps to the baseline model by letting the defender's payoffs be $n_D^1 = \frac{1}{2} - \delta_1$, $n_D^2 = \frac{1}{2} - \delta_2$, $w_D^1 = \left(\frac{1}{2} - \delta_1\right) - \phi(\delta_1) - c_D$, $w_D^2 = \left(\frac{1}{2} - \delta_2\right) - \phi(\delta_2) - c_D$, the challenger's payoffs be $n_C^1 = \frac{1}{2} + \delta_1$, $n_C^2 = \frac{1}{2} + \delta_2$, $w_C^1 = \left(\frac{1}{2} + \delta_1\right) + \phi(\delta_1) + \theta_C$, $w_C^2 = \left(\frac{1}{2} + \delta_2\right) + \phi(\delta_2) + \theta_C$, and $\theta_C \sim U\left[-\frac{1}{4}, 0\right]$.

advancement with war, and the challenger only attempts to advance if $c_C < \phi(\delta_t)$. In all other periods the defender never responds with war, and the challenger always advances. This profile produces the desired equilibrium outcomes and the challenger is best responding. So we must show that the defender doesn't wish to deviate.

Consider first a period $t < t^*$ in which the challenger attempts to advance. To get to this period the challenger must have advanced to t and the defender must have always permitted it. So this is on equilibrium path, if the defender plays his equilibrium strategy of again permitting advancement then the challenger will advance all the way to δ_t^* before advancing triggers war, and the defender's expected payoff is:

$$\left(\frac{1}{2} - \delta_t^*\right) - P\left(c_C < \phi\left(\delta_t^*\right)\right)\left(\phi\left(\delta_t^*\right) + c_D\right). \tag{C.1}$$

In words, the defender's equilibrium expected holdings are $\frac{1}{2} - \delta_t^*$, with probability $P\left(c_C < \phi\left(\delta_t^*\right)\right)$ war occurs in period t^*, and when this occurs the defender suffers the challenger's excess military advantage $\phi\left(\delta_t^*\right)$ and the cost of war c_D. If instead the defender responds with war in period t, his payoff is $\left(1 - p\left(\delta_t\right)\right) - c_D = \left(\frac{1}{2} - \delta_t - \phi\left(\delta_t\right)\right) - c_D$, which is $<$ eqn. $(C.1)$ i.f.f.

$$c_D > \frac{\left(\left(\delta_t^* + \phi\left(\delta_t^*\right)\right) - \left(\delta_t + \phi\left(\delta_t\right)\right)\right)}{\left(1 - P\left(c_C < \phi\left(\delta_t^*\right)\right)\right)} - \phi\left(\delta_t^*\right).$$

Since $\phi\left(\delta_t\right) \to 0\ \forall \delta_t$ as $Z \to \infty$, the r.h.s. approaches $\delta_t^* - \delta_t < \frac{1}{4}$ (since $\delta_t > 0$ and $\delta_t^* \leq \delta_{\bar{t}} < \frac{1}{4}$) as $Z \to \infty$. So since $c_D \geq \frac{1}{4}$ there exists a Z sufficiently large such that the inequality is satisfied for all $t < t^*$. Intuitively, we can scale down the excess military advantage function $\phi\left(\delta_t\right)$ by increasing Z sufficiently so that the calculation essentially reduces to whether the cost of war exceeds the foregone share of the landmass from allowing the challenger to advance from t all the way to t^*. This will always be true since (by assumption) the cost of war exceeds the challenger's excess holdings in the period where war occurs ($\delta_{t^*} < c_D$).

Now consider a period $t \geq \bar{t}$ in which the challenger attempts to advance. This is off path, but

we do not need beliefs about the challenger's type since if she is allowed to advance the strategies are for her to continue to advance and the defender to permit it. So if the defender allows advancement in t the challenger will eventually possess the entire landmass and the defender's payoff will be 0. If instead he responds with war his payoff is $\left(\frac{1}{2} - \delta_t - \phi\left(\delta_t\right)\right) - c_D$. Since $\delta_{\bar{t}} < \frac{1}{2} - c_D$ and $\delta_t > \frac{1}{2} - c_D$ $\forall t > \bar{t}$, for Z sufficiently large it will be optimal for the defender to respond with war in $\bar{t}$ but not in $t > \bar{t}$. In words, at $\bar{t}$ the remaining landmass just exceeds the defender's cost of war, so he will respond with war knowing that should he allow advancement he will also allow it in all future periods. For $t > \bar{t}$, the challenger is already sufficiently advanced that letting her take the remaining landmass is optimal.

Finally, consider a period $\hat{t} \in [t^*, \bar{t})$ in which the challenger attempts to advance and the defender is supposed to respond with war. The challenger already advanced in period t^* expecting to trigger war. So the defender *infers* in equilibrium that her cost $c_C < \phi\left(\delta_{t^*}\right)$, the threshold in the first period t^* in which she advanced expecting war. If she is allowed to again advance in period $\hat{t}$ to period $\hat{t}+1$, a further attempt to advance in $\hat{t}+1$ will provoke war. Anticipating this, the challenger will once again advance i.f.f. $c_C < \phi\left(\delta_{\hat{t}+1}\right)$. Recall that $\delta_{\hat{t}+1} \leq \delta_{\bar{t}} < \frac{1}{4}$ and $\phi\left(\delta_t\right)$ is increasing over $\left[0, \frac{1}{4}\right]$, so $\phi\left(\delta_{t^*}\right) < \phi\left(\delta_{\hat{t}+1}\right)$. In words, the region of the landmass is s.t. advancement makes war relatively more attractive to the challenger. So the defender can infer that a challenger who advanced to period $\hat{t}$ expecting war will again advance in period $\hat{t}+1$ even though it will trigger war for sure. So responding with war in $\hat{t}$ is optimal, since permitting advancement will only weaken the defender in the inevitable war.

We last argue that when $c_D \geq \frac{1}{2}$, equilibrium requires the challenger to occupy the entire landmass. Suppose the defender's strategy involves responding to further advancement with war with strictly positive probability in any period $t \geq 1$. This would yield utility $\left(\frac{1}{2} - \delta_t - \phi\left(\delta_t\right)\right) - c_D$ which is < 0 since $\delta_1 > 0$. If she were instead to deviate to always allowing advancement in every period, her utility would be ≥ 0; at worst the challenger's strategy will involve occupying the entire landmass

(recall that the defender holds the final decision to fight). Thus, equilibrium requires the defender to always permit advancement. Equilibrium then must also requires the challenger to attempt advancement in every period regardless of her type since it will always be permitted, and the unique equilibrium outcome is that she will occupy the entire landmass. ■

We now consider the second extension, in which the challenger makes an endogenous "demand" δ_2 of how far to advance. As in the baseline model, in this example the defender can allow a positive demand or respond with war, and if she advances the challenger can exploit her gains afterward by unilaterally initiating war.

Extension 2 (Endogenous Transgression)**.** *Consider the following $T = 2$ period game. In period 1 the challenger's excess holdings are $\delta_1 > 0$, and she can attempt to advance to some $\delta_2 \in \left[\delta_1, \frac{1}{2}\right]$ of her choosing. The defender can permit the advancement or respond with war. If he permits it, then the game proceeds to the second period, and the challenger decides whether to unilaterally initiate war or enjoy her gains. After either choice the game ends.*

The set of equilibria in this extension satisfy the following proposition.

Proposition C.2. *If $c_D \in \left[\frac{1}{4}, \frac{1}{2}\right)$ and the challenger's excess share δ_1 under the status quo is less than $\frac{1}{4} - \frac{c_D}{2}$, then there exists an equilibrium in which the defender responds to a strictly positive demand $\delta_2 \in \left(\delta_1, \frac{1}{2}\right]$, however small, with war. If $c_D \geq \frac{1}{2}$, then in any equilibrium the challenger demands the entire landmass, i.e. $\delta_2 = 1$, and it is accepted.*

Proof: We first show the desired equilibrium when $c_D \in \left[\frac{1}{4}, \frac{1}{2}\right)$; we construct an equilibrium where all demands are on-path. Challengers with cost $c_C > \phi(\delta_1)$ demand the status quo ($\delta_2^*(c_C) = \delta_1$), it is accepted, they do not initiate war in period 2, and the game ends. All challengers with cost $c_C \leq \phi(\delta_1)$ mix identically over all positive demands $\delta_2 \in \left(\delta_1, \frac{1}{2}\right]$ and the defender always responds with war. Should any such demand be accepted (off path), challengers with cost $c_C < \phi(\delta_2)$ unilaterally initiate war in the second period.

To see this is an equilibrium, consider first the defender's strategy. If he sees no demand ($\delta_2 = \delta_1$), he infers that the challenger will initiate war in the second period with probability 0 and so maintaining the status quo is optimal. Should he see a positive demand ($\delta_2 > \delta_1$), he can infer that the challenger's cost is below $\phi(\delta_1)$ but no more, since all such challengers mix identically over all positive demands. If the demand he receives satisfies $\delta_2 \in \left(\delta_1, \frac{1}{2} - \delta_1\right)$, then $\phi(\delta_1) < \phi(\delta_2)$, and since challengers with cost $c_C < \phi(\delta_2)$ will unilaterally initiate war in period 2, the probability of appeasing an already belligerent challenger by accepting such a demand is 0. Thus responding with war is optimal. If instead $\delta_2 \in \left[\frac{1}{2} - \delta_1, \frac{1}{2}\right]$, then even if allowing the demand would appease the challenger for sure the defender prefers to respond with war, since accepting such a demand will leave the defender with no more than δ_1, while responding with war leaves him with $\left(\frac{1}{2} - \delta_1 - \phi(\delta_1)\right) - c_D > \delta_1$ when $\delta_1 < \frac{1}{4} - \frac{c_D}{2}$ for sufficiently large Z.

To see that the challenger wishes to play her equilibrium strategy, first note that period 2 strategies are straightforwardly optimal since the challenger is the last mover. In period 1, any positive demand will provoke war, and all challengers with cost $c_C \leq \phi(\delta_1)$ prefer war to the status quo. So such challengers are indifferent between all positive demands and are willing to mix according to the equilibrium strategy. Finally, challengers with cost $c_C > \phi(\delta_1)$ prefer the status quo to war and so making a 0 demand $\delta_2 = \delta_1$ is optimal.

We last argue that when $c_D \geq \frac{1}{2}$, equilibrium requires the challenger to occupy the entire landmass. If she demands $\delta_2 = 1$ and it is accepted, then in the second period it is optimal to end the game with peace regardless of her type. In the first period, the defender will thus accept such a demand, since it will yield utility 0, while war yields utility $\left(\frac{1}{2} - \delta_1 - \phi(\delta_1)\right) - c_D < 0$ for $c_D \geq \frac{1}{2}$. Finally, in any equilibrium the challenger must demand $\delta_2 = 1$ in the first period; all other demands will yield strictly lower utility regardless of the defender's response, or her own anticipated strategy in the second period. ■

Proposition C.2 further demonstrates that our result is not an artifact of having a fixed size of the transgression. When the status quo division is sufficiently close to an even division, there exists equilibria in which the defender responds to *any* positive demand, however small, with war. The logic is again identical to the two-period model. At the status quo, the challenger expects the defender to respond to any positive demand with war. Hence, the defender can infer in equilibrium that a challenger who makes such a demand desires war under the status quo. Because the challenger's probability of victory $p(\delta_t)$ is such that advancements $\delta_2 \in \left(\delta_1, \frac{1}{2} - \delta_1\right)$ make war relatively more attractive, the probability of appeasing an already-belligerent challenger by permitting such an advancement is 0. Alternatively, while advancements $\delta_2 \in \left[\frac{1}{2} - \delta_1, \frac{1}{2}\right]$ have some hope of successful appeasement, they are so large that the defender prefers to suffer the cost of war.

C.2 Robustness to challenger backing down

Proposition C.3. *Consider an alternative game* Γ' *form in which the challenger can back down in the first stage if the defender resists. Whenever the deterrence equilibrium exists in the original game* Γ *it also exists in* Γ'.

Proof: In Γ' the deterrence equilibrium takes the following form; the defender always resists, the challenger is deterred unless she prefers immediate war, and when she transgresses she also fights upon resistance. Now if the defender always resists, challenger types $\theta_C \geq \bar{\theta}_C^1$ still prefer to transgress because the defender will resist and they will then proceed with war. Challenger types $\theta_C < \bar{\theta}_C^1$ cannot get away with the transgression because the defender always resists, can back down upon encountering resistance, and are therefore indifferent between transgressing and not; they are thus willing to play the required strategy of not transgressing. Upon observing a transgression the defender therefore continues to infer that the challenger is of type $\theta_C \geq \bar{\theta}_C^1$, and in this case resisting is equivalent to unilaterally initiating war himself; his incentives and inferences are unchanged and

he is therefore willing to carry out his equilibrium strategy. ■

C.3 Game with interdependent war values

Suppose that both players' payoffs in the event of war depend on the challenger's type $\theta_C \in \Theta \subset \mathbb{R}$ that is unknown to the defender but known to the challenger, where Θ is an interval and θ_C has a prior distribution $f(\theta_C)$ with full support over Θ. The challenger's type is therefore to be interpreted as a state of the world that affects both players' payoffs over which the challenger has private information. Our notation and assumptions for the challenger's payoffs are unchanged. For the defender, we now express the dependence of his war payoff on the challenger's type using $w_D^t(\theta_C)$, and make the following slightly-modified assumptions.

1. For all challenger types, allowing the transgression makes the defender strictly worse off in both peace $\left(n_D^2 < n_D^1\right)$ and war $\left(w_D^2(\theta_C) < w_D^1(\theta_C)\ \ \forall \theta_C\right)$.

2. For all challenger types, allowing the transgression is strictly better than responding with war if the challenger will subsequently choose peace $\left(n_D^2 > w_D^1(\theta_C)\ \ \forall \theta_C\right)$.

Note that our defender assumptions jointly imply that the defender strictly prefers peace to war in each t for every type of challenger. Moreover, conditional on defender assumptions $(1) - (2)$, any arbitrary dependence of the defender's war payoff $w_D^t(\theta_C)$ on the challenger's type can be accommodated. However, it is natural to assume that $w_D^t(\theta_C)$ is weakly decreasing in θ_C, i.e., a more belligerent challenger means a weaker defender. Our setup is not completely without loss of generality because it cannot capture when the challenger is privately informed about factors affecting the defender's war payoffs but not her own; however, it is sufficiently general to capture private information about the probability of victory.

Challenger Incentives In the second period, the challenger transgresses i.f.f. $\theta_C \geq \bar{\theta}_C^2$. In the first period, challengers of type $\theta_C \geq \bar{\theta}_C^1$ always transgress. Challengers of type $\theta_C < \bar{\theta}_C^1$ transgress i.f.f.,

$$\alpha \cdot w_C^1(\theta_C) + (1-\alpha) \cdot \max\left\{n_C^2, w_C^2(\theta_C)\right\} \geq n_C^1.$$

For each such type, there exists a unique interior probability $\hat{\alpha}(\theta_C)$ that would make them indifferent between transgressing and not, and given that probability the challenger would play a cutpoint strategy at θ_C. It is simple to verify that for $\theta_C \leq \bar{\theta}_C^1$, $\hat{\alpha}(\theta_C)$ is always well defined, strictly increasing in θ_C, strictly interior to $(0,1)$, and $\hat{\alpha}\left(\bar{\theta}_C^1\right) = 1$.

Defender's Incentives Suppose that the challenger uses a threshold for transgressing equal to $\hat{\theta}_C$. Then upon observing a transgression, the defender's payoff from war is

$$\int_{\hat{\theta}_C}^{\infty} w_D^1(\theta_C) \frac{f(\theta_C)}{1-F\left(\hat{\theta}_C\right)} d\theta_C$$

and from appeasement is,

$$\int_{\hat{\theta}_C}^{\max\left\{\hat{\theta}_C, \bar{\theta}_C^2\right\}} n_D^2 \frac{f(\theta_C)}{1-F\left(\hat{\theta}_C\right)} d\theta_C + \int_{\max\left\{\hat{\theta}_C, \bar{\theta}_C^2\right\}}^{\infty} w_D^2(\theta_C) \cdot \frac{f(\theta_C)}{1-F\left(\hat{\theta}_C\right)} d\theta_C.$$

Hence she will prefer to respond to the transgression with war i.f.f.

$$\int_{\max\left\{\hat{\theta}_C, \bar{\theta}_C^2\right\}}^{\infty} \left(w_D^1(\theta_C) - w_D^2(\theta_C)\right) \cdot \frac{f(\theta_C)}{1-F\left(\hat{\theta}_C\right)} \geq \int_{\hat{\theta}_C}^{\max\left\{\hat{\theta}_C, \bar{\theta}_C^2\right\}} \left(n_D^2 - w_D^1(\theta_C)\right) \frac{f(\theta_C)}{1-F\left(\hat{\theta}_C\right)} d\theta_C$$

Now it is straightforward to show that the condition above is satisfied i.f.f.

$$\bar{\beta}\left(\hat{\theta}_C\right) \leq P\left(\theta \geq \bar{\theta}_C^2 \,|\, \theta \geq \hat{\theta}_C\right), \tag{C.2}$$

where

$$\bar{\beta}\left(\hat{\theta}_C\right) = \frac{n_D^2 - E\left[w_D^1(\theta_C) \mid \theta_C \in \left[\hat{\theta}_C, \bar{\theta}_C^2\right]\right]}{\left(n_D^2 - E\left[w_D^1(\theta_C) \mid \theta_C \in \left[\hat{\theta}_C, \bar{\theta}_C^2\right]\right]\right) + E\left[w_D^1(\theta_C) - w_D^2(\theta_C) \mid \theta_C \geq \bar{\theta}_C^2\right]} \tag{C.3}$$

Intuitively, $n_D^2 - E\left[w_D^1(\theta_C) \mid \theta_C \in \left[\hat{\theta}_C, \bar{\theta}_C^2\right]\right]$ is the benefit from appeasement conditional on the challenger being appeasable. Similarly, $E\left[w_D^1(\theta_C) - w_D^2(\theta_C) \mid \theta_C \geq \bar{\theta}_C^2\right]$ is the benefit from preemptive war conditional on the challenger being unappeasable. Finally, as before $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \hat{\theta}_C\right)$ is the interim probability that the challenger is unappeasable.

Now note the following. First, $\bar{\beta}\left(\hat{\theta}_C\right)$ is strictly interior to $[0,1]$ for any value of $\hat{\theta}_C$ by our payoff assumptions, since appeasement is beneficial when it is possible $\left(\theta_C \in \left[\hat{\theta}_C, \bar{\theta}_C^2\right]\right)$ and war early is better than war later when it is not $\left(\theta_C > \bar{\theta}_C^2\right)$. Second, $\bar{\beta}\left(\hat{\theta}_C\right)$ is weakly increasing in $\hat{\theta}_C$ in the natural case where a belligerent challenger is "bad news" for the defender (i.e. $w_D^t(\theta_C)$ is decreasing in θ_C) since then $E\left[w_D^1(\theta_C) \mid \theta_C \in \left[\hat{\theta}_C, \bar{\theta}_C^2\right]\right]$ is decreasing in $\hat{\theta}_C$. Third and as in the baseline model, $P\left(\theta \geq \bar{\theta}_C^2 \mid \theta \geq \hat{\theta}_C\right)$ is increasing in $\hat{\theta}_C$ – that is, the defender's interim assessment that appeasement will be ineffective is higher when the challenger uses a higher threshold for transgressing.

Equilibrium Characterization Applying the analysis above, we now have the following complete equilibrium characterization.

Proposition C.4. *Equilibria of the model with interdependent values are as follows.*

- *The deterrence equilibrium exists i.f.f.*

$$\bar{\beta}\left(\bar{\theta}_C^1\right) \leq P\left(\theta \geq \bar{\theta}_C^2 \mid \theta \geq \bar{\theta}_C^1\right)$$

- *The no deterrence equilibrium exists i.f.f.*

$$P\left(\theta_C \geq \bar{\theta}_C^2\right) \leq \bar{\beta}\left(-\infty\right)$$

- *A mixed strategy equilibrium in which the challenger uses threshold* $\hat{\theta}_C^* < \min\left\{\bar{\theta}_C^1, \bar{\theta}_C^2\right\}$ *exists i.f.f*

$$\bar{\beta}\left(\hat{\theta}_C^*\right) = P\left(\theta \geq \bar{\theta}_C^2 \,|\, \theta \geq \hat{\theta}_C^*\right)$$

In the equilibrium, the defender responds to the transgression with war with probability $\alpha^* = \hat{\alpha}\left(\hat{\theta}_C^*\right)$.

The most important observation from the above characterization is the following: because $\bar{\beta}\left(\hat{\theta}_C\right)$ is interior for all $\hat{\theta}_C$ (meaning that war sooner is better than war later), our basic insight holds unaltered. When appeasement is ineffective $\left(\bar{\theta}_C^2 \leq \bar{\theta}_C^1\right)$, the deterrence equilibrium exists for all distributions over the challenger's type θ_C and functions $w_D^t\left(\theta_C\right)$ mapping the challenger's type into the defender's payoff from war that satisfy the initial assumptions. Thus, Corollaries 1 and 3 continue to hold unaltered with interdependent values.

Other more subtle patterns of equilibria can occur with interdependent values. Because $\bar{\beta}\left(\hat{\theta}_C\right)$ can be steeply increasing in $\hat{\theta}_C$ rather than constant, it is no longer the case that the mixed strategy equilibrium can only exist when both pure strategy equilibria exist. Many different scenarios can occur, including an odd number of mixed strategy equilibria combined with an even number of pure strategy equilibria (including none), and a single pure strategy equilibrium combined with an even number of mixed strategy equilibria.

Intuitively, the reason for this multiplicity of equilibria is that a higher threshold for transgressing by the challenger has two countervailing effects. First, it makes the defender *less* willing to appease because her interim assessment of the probability that the challenger is unappeasable is higher.

Second, it makes the challenger *more* willing to appease because inferring the challenger is a higher type also means that war is worse, making appeasement more attractive if it can be effective. These countervailing effects can then generate multiple equilibria: with higher thresholds, the defender can find appeasement less likely to be effective, but simultaneously more desirable if it would be effective.

C.4 Game with two-sided uncertainty

The defender is now assumed to have a type θ_D upon which his war payoffs in each period depend, so we write $w_D^t(\theta_D)$ to express this dependence. We maintain the assumption that payoffs in peace for both players are fixed and common knowledge, and make new assumptions on the defender's type that mirror those of the challenger. Specifically, θ_D also belongs to an interval, has some prior distribution $g(\theta_D)$ with full support, and is distributed independently of θ_C. Thus, war values are private and each side's uncertainty may be interpreted as about the opponent's cost of war. We modify the assumptions the defender's payoffs as follows:

1. For all defender types, allowing the transgression makes the defender strictly worse off in both peace $\left(n_D^2 < n_D^1\right)$ and war $\left(w_D^2(\theta_D) < w_D^1(\theta_D) \ \ \forall \theta_D\right)$.

2. In each period t the defender's war payoff $w_D^t(\theta_D)$ is continuous and strictly increasing in θ_D. In addition, there exists a unique defender type $\bar{\theta}_D^t$ that is indifferent between peace and war in period t.

3. The benefit $w_D^1(\theta_D) - w_D^2(\theta_D) > 0$ of war sooner vs. war is weakly increasing in the defender's type.

The first assumption extends the properties of the transgression to a setting where the defender's payoffs can vary, and the second mirrors the assumptions made on the challenger's type. Importantly, it implies that with strictly positive probability the defender's threat is "inherently" credible in that

he is willing to go to war solely to prevent the transgression. Formally, for both players let $\bar{\theta}_i^{s,t}$ denote a player indifferent between peace in period s and war in period t – since $n_D^2 < n_D^1$ we have $\bar{\theta}_D^{2,1} < \bar{\theta}_D^{1,1}$ and types in between are willing to fight a war over the transgression.

The third assumption ensures that types who are overall more belligerent are also weakly more willing to go to war for preemptive reasons, and is necessary for the existence of cutpoint strategies. Finally, since the defender may unilaterally wish to initiate war in both periods, we augment the first period with a final stage in which the defender can start a war even if the challenger chooses not to transgress. It is unnecessary to augment the second period with a similar stage because any defender type who would unilaterally initiate war in the second stage would also initiate war in the first stage and end the game.

Challenger Incentives Challenger incentives are identical to the game with interdependent war values except for the following distinction – because the defender may now be of a type $\theta_D \geq \bar{\theta}_D^1$ who would start a war whether or not the challenger attempts to transgresses, α now denotes the probability that transgressing would *provoke* an otherwise peaceful challenger to start a war. If the defender uses a cutpoint strategy $\hat{\theta}_D \leq \bar{\theta}_D^1$ for responding to the transgression, then in equilibrium $\alpha = \frac{G\left(\bar{\theta}_D^1\right) - G\left(\hat{\theta}_D\right)}{G\left(\bar{\theta}_D^1\right)}$.

Defender's Incentives The defender's war payoffs now depend on her type θ_D; moreover, because types are independent the threshold $\hat{\theta}_C$ that the challenger uses for transgressing only affects her payoffs through the interim assessment β that the challenger would initiate war after being allowed to transgress. He therefore prefers to respond to the transgression with war when $\beta \geq \bar{\beta}\left(\theta_D\right)$, where

$$\bar{\beta}\left(\theta_D\right) = \frac{n_D^2 - w_D^1\left(\theta_D\right)}{\left(n_D^2 - w_D^1\left(\theta_D\right)\right) + \left(w_D^1\left(\theta_D\right) - w_D^2\left(\theta_D\right)\right)}. \tag{C.4}$$

It is simple to verify that for $\theta_D \in [0, \bar{\theta}_D^{2,1})$ (where $\bar{\theta}_D^{2,1}$ is the defender type indifferent between

immediate war and successful appeasement) the function $\bar{\beta}(\theta_D)$ is strictly interior to $[0,1]$ and decreasing (by assumption 3). The latter property ensures that the defender always plays a cutpoint strategy, and we can therefore also work with the inverse function $\bar{\theta}_D(\beta) = \bar{\beta}^{-1}(\beta)$ denoting the defender type indifferent between appeasement and war when his interim assessment is β.

Equilibrium Characterization

Proposition C.5. *Equilibria of the model with two-sided uncertainty are as follows.*

- *The deterrence equilibrium exists i.f.f.*

$$\bar{\beta}(-\infty) \leq P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta \geq \bar{\theta}_C^1\right)$$

- *The no deterrence equilibrium exists i.f.f.*

$$P\left(\theta_D \in \left[\bar{\theta}_D\left(P\left(\theta_C \geq \bar{\theta}_C^2\right)\right), \bar{\theta}_D^1\right] \,|\, \theta_D \leq \bar{\theta}_D^1\right) \leq \hat{\alpha}(-\infty)$$

- *An interior equilibrium with challenger threshold $\hat{\theta}_C^* < \min\left\{\bar{\theta}_C^1, \bar{\theta}_C^2\right\}$ exists i.f.f.*

$$P\left(\theta_D \in \left[\bar{\theta}_D\left(P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \hat{\theta}_C^*\right)\right), \bar{\theta}_D^1\right] \,|\, \theta_D \leq \bar{\theta}_D^1\right) = \hat{\alpha}\left(\hat{\theta}_C^*\right)$$

or equivalently

$$\frac{G\left(\bar{\theta}_D\left(\frac{1-F\left(\bar{\theta}_C^2\right)}{1-F\left(\hat{\theta}_C^*\right)}\right)\right) - G\left(\bar{\theta}_D^1\right)}{G\left(\bar{\theta}_D^1\right)} = \hat{\alpha}\left(\hat{\theta}_C^*\right)$$

In the equilibrium, the challenger transgresses when $\theta_C \geq \hat{\theta}_C^*$ and the defender responds with war i.f.f. $\theta_D \geq \bar{\theta}_D\left(P\left(\theta_C \geq \bar{\theta}_C^2 \,|\, \theta_C \geq \hat{\theta}_C^*\right)\right) = \bar{\theta}_D\left(\frac{1-F\left(\bar{\theta}_C^2\right)}{1-F\left(\hat{\theta}_C^*\right)}\right)$.

Again, the most important observation from the above characterization is that because $\bar{\beta}\left(\hat{\theta}_C\right)$

is interior for all $\hat{\theta}_C$ (meaning that war sooner is better than war later), our basic insight again holds unaltered. When appeasement is ineffective $\left(\bar{\theta}_C^2 \leq \bar{\theta}_C^1\right)$, the deterrence equilibrium exists for all distributions over the challenger's type θ_C and defender's type θ_D that satisfy the initial assumptions, and Corollaries 1 and 3 hold unaltered.

As with interdependent war values other more subtle patterns of equilibria can also occur. Intuitively, the reason is that deterrence begets deterrence – a higher threshold for transgressing (greater $\hat{\theta}_C$) generates a higher interim assessment $\frac{1-F\left(\bar{\theta}_C^2\right)}{1-F\left(\hat{\theta}_C^*\right)}$ that the challenger is unappeasable, generating a lower threshold $\bar{\theta}_D\left(\frac{1-F\left(\bar{\theta}_C^2\right)}{1-F\left(\hat{\theta}_C^*\right)}\right)$ for the defender to respond with war, a higher probability $\frac{G\left(\bar{\theta}_D\left(\frac{1-F\left(\bar{\theta}_C^2\right)}{1-F\left(\hat{\theta}_C^*\right)}\right)\right)-G\left(\bar{\theta}_D^1\right)}{G\left(\bar{\theta}_D^1\right)}$ that the defender will be provoked by an attempted transgression, and thus more deterrence. Under some conditions this dynamic can set off a "deterrence spiral" where the challenger is very unlikely to be unappeasable ex-ante yet the deterrence equilibrium is unique – a sufficient condition for this occurring is the standard condition that the "no deterrence" equilibrium be unstable and the slope of the challenger's best response function

$$\hat{\alpha}^{-1}\left(\frac{G\left(\bar{\theta}_D\left(\frac{1-F\left(\bar{\theta}_C^2\right)}{1-F\left(\hat{\theta}_C^*\right)}\right)\right)-G\left(\bar{\theta}_D^1\right)}{G\left(\bar{\theta}_D^1\right)}\right)$$

be greater than 1 (where $\hat{\alpha}^{-1}(\alpha)$ denotes the well-defined inverse of $\hat{\alpha}(\theta_C)$).

C.5 Game with uncertainty about transgression's military consequences

We now consider robustness to ex-ante uncertainty about the military consequences of the transgression. Suppose that the challenger's second period war payoff is $w_C^2(\theta) - \sigma\varepsilon$, where ε is a symmetric random variable with mean 0, variance 1, and atomless full support (so that $\sigma\varepsilon$ has variance σ^2). Let $H(\varepsilon)$ denote the CDF describing the distribution of ε. The realization of ε is initially unknown to both players, and then publicly revealed following a successful transgression. This captures the idea

that the *realized* military benefits $\left(w_C^2(\theta_C) - \sigma\varepsilon\right) - w_C^1(\theta_C)$ of the transgression to the challenger may be more or less than the initial expected value $w_C^2(\theta_C) - w_C^1(\theta_C) = \delta_C^m(\theta_C)$.

Challenger Incentives In the second period, the challenger initiates war i.f.f. $\varepsilon \leq \frac{w_C^2(\theta_C) - n_C^2}{\sigma}$. So from an ex-ante perspective, a challenger with type θ_C will initiate war in the second period with probability $H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma}\right)$.

In the first period it is easily verified that challengers of type $\theta_C \geq \bar{\theta}_C^1$ always transgress. Challengers of type $\theta_C < \bar{\theta}_C^1$ transgress i.f.f.,

$$\alpha \cdot w_C^1(\theta_C) + (1-\alpha)\int \max\left\{n_C^2, w_C^2(\theta_C) - \sigma\varepsilon\right\} h(\varepsilon)\, d\varepsilon \geq n_C^1$$

It is also easily shown that for each such type, there exists a unique interior probability $\hat{\alpha}(\theta_C)$ that would make them indifferent between transgressing and not, and given that probability the challenger would play a cutpoint strategy at θ_C. Finally, it is simple to verify that for $\theta_C \leq \bar{\theta}_C^1$, $\hat{\alpha}(\theta_C)$ is always well defined, strictly increasing in θ_C, strictly interior to $(0,1)$, and $\hat{\alpha}\left(\bar{\theta}_C^1\right) = 1$.

Defender Incentives Defender incentives are as in the baseline model; he prefers to respond to an attempted transgression with war i.f.f. his interim belief β is $> \bar{\beta} = \frac{n_D^2 - w_D^1}{(n_D^2 - w_D^1) + (w_D^1 - w_D^2)}$. It is also helpful to denote $\bar{\theta}_C^\sigma$ as the challenger type against whom the defender would be exactly indifferent to responding to the transgression with war; i.e., $H\left(\frac{w_C^2(\bar{\theta}_C^\sigma) - n_C^2}{\sigma}\right) = \bar{\beta}$. Note that no such type exists in the baseline model because the challenger's probability of initiating war after transgressing increases discontinuously from 0 to 1 when θ_C crosses $\bar{\theta}_C^2$.

We must also characterize the defender's interim beliefs about the probability a challenger who transgressed will initiate war in the second period when she is *uncertain* of the challenger's type; given the analysis in the preceding section we may restrict attention to cutpoint strategies by the challenger. When the challenger plays a cutpoint strategy of transgressing i.f.f. $\theta_C \geq \hat{\theta}_C$, it is straightforward

to see that the defender's interim assessment of this probability is $E_{\theta_C}\left[H\left(\frac{w_C^2(\theta_C)-n_C^2}{\sigma}\right) \;\middle|\; \theta_C \geq \hat{\theta}_C\right]$. It is also easy to show that the derivative of this probability w.r.t. the challenger's cutpoint $\hat{\theta}_C$ is

$$\frac{f\left(\hat{\theta}_C\right)}{1-F\left(\hat{\theta}_C\right)} \int\limits_{\theta_C \geq \hat{\theta}_C} \left(H\left(\frac{w_C^2\left(\theta_C\right)-n_C^2}{\sigma}\right) - H\left(\frac{w_C^2\left(\hat{\theta}_C\right)-n_C^2}{\sigma}\right)\right) \frac{f\left(\theta_C\right)}{1-F\left(\hat{\theta}_C\right)} d\theta_C.$$

This is > 0 since $\frac{w_C^2(\theta_C)-n_C^2}{\sigma}$ is strictly increasing in θ_C. The defender's interim assessment is thus smoothly increasing in the challenger's cutpoint $\hat{\theta}_C$, approaches the prior $E_{\theta_C}\left[H\left(\frac{w_C^2(\theta_C)-n_C^2}{\sigma}\right)\right]$ as $\theta_C \to \inf\Theta$, and approaches $\lim\limits_{\theta_C \to \infty} H\left(\frac{w_C^2(\theta_C)-n_C^2}{\sigma}\right)$ as $\theta_C \to \sup\Theta$.

Equilibrium Characterization Using the above characterizations, equilibrium is as follows.

Proposition C.6. *With uncertainty about the transgression's military value to the challenger,*

1. *there exists a* ***no deterrence equilibrium,*** *in which the challenger always transgresses, and she is always permitted to do so, i.f.f.*

$$\bar{\beta} \geq E_{\theta_C}\left[H\left(\frac{w_C^2\left(\theta_C\right)-n_C^2}{\sigma}\right)\right]$$

2. *there exists a* ***deterrence equilibrium****, in which (i) the defender always responds to the transgression with war, (ii) all types $\theta_C < \bar{\theta}_C^1$ who do not initially prefer war are deterred, and (iii) the probability of deterrence is $P\left(\theta_1 < \bar{\theta}_C^1\right)$, i.f.f.*

$$\bar{\beta} \leq E_{\theta_C}\left[H\left(\frac{w_C^2\left(\theta_C\right)-n_C^2}{\sigma}\right) \;\middle|\; \theta_C \geq \bar{\theta}_C^1\right]$$

3. *a mixed strategy equilibrium exists i.f.f. both the no deterrence and deterrence equilibria exist,*

and is uniquely characterized by both a challenger cutpoint $\hat{\theta}_C^*$ satisfying

$$\bar{\beta} = E_{\theta_C}\left[H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma}\right) \;\middle|\; \theta_C \geq \hat{\theta}_C^*\right]$$

and a defender probability of responding to the transgression with war equal to $\alpha^* = \hat{\alpha}\left(\hat{\theta}_C^*\right)$. When the mixed strategy equilibrium exists the defender is best off in the deterrence equilibrium.

To summarize, equilibrium incentives and conditions are effectively identical to the baseline model, except that the idiosyncratic shock $\sigma\varepsilon$ "smooths out" the defender's interim assessment of the probability that the challenger will initiate war if allowed to transgress. The proof is simply the assembly of the preceding analysis except for the payoff dominance of the deterrence equilibrium when there are multiple equilibria – this is straightforward to show using an identical argument as in Proposition 1 but with the defender's interim assessments suitably modified.

Results We now restate analogs to our main results in this variant of the model.

Proposition C.7. *With uncertainty about the transgression's military value to the challenger,*

1. *if $\bar{\theta}_C^\sigma \leq \bar{\theta}_C^1$ then the deterrence equilibrium exists.*

2. *if absent uncertainty the deterrence equilibrium exists strictly $\left(\bar{\beta} > P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right)\right)$, then it also exists with uncertainty for sufficiently small σ. Consequently, if $\bar{\theta}_C^2 < \bar{\theta}_C^1 \iff \delta_C^m\left(\bar{\theta}_C^1\right) > \delta_C^d$ then the deterrence equilibrium exists for sufficiently small σ.*

3. *under the assumptions stated in Proposition 3, the probability that deterrence is successful is increasing in $\delta_C^m - \delta_C^d$.*

4. *if the deterrence equilibrium prevails whenever it exists, then the probability of deterrence would decrease if the defender knew the challenger's type. In addition, the defender is better off not*

knowing the challenger's type i.f.f. either $\bar{\theta}_C^{\sigma} \leq \bar{\theta}_C^1$*, or* $\bar{\theta}_C^{\sigma} > \bar{\theta}_C^1$ *and*

$$\int_{\bar{\theta}_C^1}^{\bar{\theta}_C^{\sigma}} \left(\bar{\beta} - H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma} \right) \right) f(\theta_C)\, d\theta_C \leq$$

$$\min \left\{ \int_0^{\bar{\theta}_C^1} \left(\bar{\beta} \left(\frac{n_D^1 - n_D^2}{n_D^2 - w_D^1} \right) + H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma} \right) \right) f(\theta_C)\, d\theta_C, \int_{\bar{\theta}_C^{\sigma}}^{\infty} \left(H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma} \right) - \bar{\beta} \right) f(\theta_C)\, d\theta_C \right\}$$

Proposition C.7 shows that analogues of our main results hold in this variant of the model; for the purposes of exposition the proof is deferred to the end of this section. The reason is simple. Introducing uncertainty about the consequences of the transgression (captured by the size of σ) weakens the extent to which future belligerence can be inferred from present belligerence. However, it does not change the fact that a higher relative military gain strengthens this inference, nor the defender's fundamental incentives. Consequently, our results are not knife edge – introducing a little bit of uncertainty about the transgression's consequences does not perturb them, but with enough uncertainty the conditions for deterrence may fail to hold.

Specifically, the proposition first provides an analogue condition $\left(\bar{\theta}_C^{\sigma} \leq \bar{\theta}_C^1\right)$ to the existence condition $\left(\bar{\theta}_C^2 \leq \bar{\theta}_C^1\right)$ for the deterrence equilibrium in the baseline model; that the challenger type at which the defender switches from preferring appeasement to preferring war be weakly less than the challenger type indifferent between peace and war in the first period. Absent uncertainty about the transgression's consequences this type is exactly $\bar{\theta}_C^2$ – that is, the challenger type indifferent between peace and war in the second period. The reason is that absent uncertainty this is challenger type at which appeasement switches from being an effective to ineffective strategy. With uncertainty however, this type is a more complex function of the uncertainty and underlying payoffs.

The second part of the proposition states that introducing a little bit of uncertainty about the transgression's consequences does not perturb the deterrence equilibrium; that is, if deterrence works without uncertainty, then it will also work with a little bit of uncertainty. This implies that deterrence

always works when the transgression's *expected* military value exceeds its direct value ($\delta_C^m - \delta_C^d > 0$) even when a little bit of uncertainty is introduced. Relatedly, the third part of the proposition states that with uncertainty it remains true that the probability of successful deterrence is increasing in $\delta_C^m - \delta_C^d$. Both with and without this uncertainty, a challenger who prefers war to peace in the first period is ex-ante more likely to prefer war after transgressing the larger is this difference.

The last part of the proposition states that knowing the challenger's type still reduces the probability of deterrence. With such knowledge the challenger can only successfully deter when $\bar{\theta}_C^\sigma < \bar{\theta}_C^1$ (where $\bar{\theta}_C^\sigma = \bar{\theta}_C^2$ when $\sigma = 0$), and moreover can only deters types $\theta_C \in \left[\bar{\theta}_C^\sigma, \bar{\theta}_C^1\right]$. However, when she lacks this knowledge, then under these conditions the deterrence equilibrium exists, and in it she deters all types $\theta_C \leq \bar{\theta}_C^1$. Finally, the proposition states analogous conditions to those in the baseline model for the defender to be better off lacking knowledge of the challenger's type; this is the conjunction of the deterrence equilibrium existing, and the benefits of deterring types $\theta_C \leq \bar{\theta}_C^1$ outweighing the costs of fighting wars against types $\theta_C \in \left[\bar{\theta}_C^1, \bar{\theta}_C^\sigma\right]$ that she'd prefer to appease.

Proof of Proposition C.7 **(Part 1)** Observe that

$$\bar{\beta} = H\left(\frac{w_C^2\left(\bar{\theta}_C^\sigma\right) - n_C^2}{\sigma}\right) \leq E_{\theta_C}\left[H\left(\frac{w_C^2\left(\theta_C\right) - n_C^2}{\sigma}\right) \;\middle|\; \theta_C \geq \bar{\theta}_C^\sigma\right] \leq E_{\theta_C}\left[H\left(\frac{w_C^2\left(\theta_C\right) - n_C^2}{\sigma}\right) \;\middle|\; \theta_C \geq \bar{\theta}_C^1\right],$$

so the deterrence equilibrium exists. The first equality follows from the definition of $\bar{\theta}_C^\sigma$ and the remaining inequalities from previous arguments.

(Part 2) Comparing the conditions for the deterrence equilibrium in Propositions 1 and C.6, it suffices to show that $\lim_{\sigma \to 0}\left\{E_{\theta_C}\left[H\left(\frac{w_C^2(\theta_C) - n_C^2}{\sigma}\right) \;\middle|\; \theta_C \geq \bar{\theta}_C^1\right]\right\} = P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right)$. Observe that $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right) = \int\limits_{\theta_C \geq \bar{\theta}_C^1} \mathbf{1}_{\theta_C \geq \bar{\theta}_C^2} \cdot \frac{f(\theta_C)}{1 - F\left(\bar{\theta}_C^1\right)} d\theta_C$ and

$$E_{\theta_C}\left[H\left(\frac{w_C^2\left(\theta_C\right) - n_C^2}{\sigma}\right) \;\middle|\; \theta_C \geq \bar{\theta}_C^1\right] = \int\limits_{\theta_C \geq \bar{\theta}_C^1} H\left(\frac{w_C^2\left(\theta_C\right) - n_C^2}{\sigma}\right) \frac{f\left(\theta_C\right)}{1 - F\left(\bar{\theta}_C^1\right)} d\theta_C$$

The desired result then follows from the dominated convergence theorem because $H\left(\frac{w_C^2(\theta_C)-n_C^2}{\sigma}\right) \leq 1$ $\forall\,(\theta_C,\sigma)$ and converges pointwise almost everywhere to $\mathbf{1}_{\theta_C \geq \bar{\theta}_C^2}$ (since $\lim_{\sigma\to 0}\left\{H\left(\frac{w_C^2(\theta_C)-n_C^2}{\sigma}\right)\right\} = 0$ for $w_C^2(\theta_C) < n_C^2 \iff \theta_C < \bar{\theta}_C^2$ and $\lim_{\sigma\to 0}\left\{H\left(\frac{w_C^2(\theta_C)-n_C^2}{\sigma}\right)\right\} = 1$ for $w_C^2(\theta_C) > n_C^2 \iff \theta_C > \bar{\theta}_C^2$). Last the stated implications of $\bar{\theta}_C^2 \leq \bar{\theta}_C^1$ follow from the above and $P\left(\theta_C \geq \bar{\theta}_C^2 \mid \theta_C \geq \bar{\theta}_C^1\right) = 1 > \bar{\beta}$.

(Part 2) Given the assumptions in the proposition and holding $\bar{\theta}_C^1$ (i.e. the challenger's first period payoffs) fixed, the probability of deterrence is 0 if $E_{\theta_C}\left[H\left(\frac{\left(w_C^1(\theta_C)-n_C^1\right)}{\sigma} + \left(\frac{\delta_C^m-\delta_C^d}{\sigma}\right)\right) \,\middle|\, \theta_C \geq \bar{\theta}_C^1\right] < \bar{\beta}$ and $P\left(\theta_C \leq \bar{\theta}_C^1\right)$ otherwise. The quantity on the l.h.s. is self-evidently increasing in $\delta_C^m - \delta_C^d$; hence the result is shown.

(Part 3) If the defender knew the challengers type, he would respond with war i.f.f. $H\left(\frac{w_C^2(\theta_C)-n_C^2}{\sigma}\right) \geq \bar{\beta}$, implying that the challenger would be deterred i.f.f. $\theta_C \in \left(\bar{\theta}_C^\sigma, \bar{\theta}_C^1\right)$. The probability of deterrence would therefore be $P\left(\theta_1 < \bar{\theta}_C^1, \theta_1 \geq \bar{\theta}_C^\sigma\right)$. Now suppose first that the deterrence equilibrium exists when the challenger's type is unknown; then the probability of deterrence is $P\left(\theta_1 < \bar{\theta}_C^1\right)$, which is $>$ the probability of deterrence $P\left(\theta_1 < \bar{\theta}_C^1, \theta_1 \geq \bar{\theta}_C^\sigma\right)$ when the challenger's type is known. Next suppose that the deterrence equilibrium does not exist when the challenger's type is unknown, so that the probability of deterrence is 0. Then by Part 1 we must have $\bar{\theta}_C^\sigma > \bar{\theta}_C^1$, which implies that the probability of deterrence is also 0 when the challenger's type is known.

Next we consider when the defender is better off not knowing the challenger's type. This is clearly the case when $\bar{\theta}_C^\sigma \leq \bar{\theta}_C^1$; types $< \bar{\theta}_C^\sigma$ are deterred when they otherwise would not be, and for all other types the outcome is identical. So suppose that $\bar{\theta}_C^1 < \bar{\theta}_C^\sigma$, and henceforth for notational simplicity denote $H\left(\frac{w_C^2(\theta_C)-n_C^2}{\sigma}\right) = p(\theta_C)$. First observe that a *necessary* condition for the defender to be better off not knowing is that the deterrence equilibrium exists; if it does not, then with uncertainty the challenger will always transgress and the defender will always appease, but lacking uncertainty the defender can identify the most warlike types $\theta_C \geq \bar{\theta}_C^\sigma$ and fight them early. We can write the

existence condition in Propositions C.6.2 in a more usable form as:

$$\begin{aligned} & \int_{\bar{\theta}_C^1}^{\infty} \left(\left(\left(1 - p\left(\theta_C\right)\right) n_D^2 + p\left(\theta_C\right) w_D^2 \right) - w_D^1 \right) \frac{f\left(\theta_C\right)}{1 - P\left(\bar{\theta}_C^1\right)} d\theta_C \leq 0 \\ \iff \quad & \int_{\bar{\theta}_C^1}^{\bar{\theta}_C^{\sigma}} \left(\bar{\beta} - p\left(\theta_C\right)\right) f\left(\theta_C\right) d\theta_C \leq \int_{\bar{\theta}_C^{\sigma}}^{\infty} \left(p\left(\theta_C\right) - \bar{\beta}\right) f\left(\theta_C\right) d\theta_C \end{aligned} \tag{C.5}$$

where the second line comes from observing that the ex-ante net benefit of appeasement $\left(1 - p\left(\theta_C\right)\right) n_D^2 + p\left(\theta_C\right) w_D^2 - w_D^1$ can be rewritten as $\left(n_D^2 - w_D^2\right)\left(\bar{\beta} - p\left(\theta_C\right)\right)$.

Last we must characterize when the defender prefers not knowing and the deterrence equilibrium to knowing the challenger's type. Comparing the two scenarios, outcomes are the same when $\bar{\theta}_C > \bar{\theta}_C^{\sigma}$ (war in the first period), there is a deterrence benefit of not knowing when $\bar{\theta}_C \leq \bar{\theta}_C^1$, and there is a cost of fighting early wars when $\theta_C \in \left[\bar{\theta}_C^1, \bar{\theta}_C^{\sigma}\right]$. To prefer not knowing therefore requires that:

$$\int_{\bar{\theta}_C^1}^{\bar{\theta}_C^{\sigma}} \left(\left(\left(1 - p\left(\theta_C\right)\right) n_D^2 + p\left(\theta_C\right) w_D^2 \right) - w_D^1 \right) f\left(\theta_C\right) d\theta_C \leq \int_0^{\bar{\theta}_C^1} \left(n_D^1 - \left(\left(1 - p\left(\theta_C\right)\right) n_D^2 + p\left(\theta_C\right) w_D^2 \right) \right) f\left(\theta_C\right) d\theta_C$$

$$\iff \int_{\bar{\theta}_C^1}^{\bar{\theta}_C^{\sigma}} \left(\bar{\beta} - p\left(\theta_C\right)\right) f\left(\theta_C\right) d\theta_C \leq \int_0^{\bar{\theta}_C^1} \left(\bar{\beta} \left(\frac{n_D^1 - n_D^2}{n_D^2 - w_D^1} \right) + p\left(\theta_C\right) \right) f\left(\theta_C\right) d\theta_C, \tag{C.6}$$

where the second line comes from using the previously observed equality and observing that $\left(\frac{n_D^1 - n_D^2}{n_D^2 - w_D^2}\right) = \bar{\beta}\left(\frac{n_D^1 - n_D^2}{n_D^2 - w_D^1}\right)$. The conjunction of equations C.5 and C.6 then yields the condition in the Proposition.

C.6 A challenger who can send a costly signal

In this section we consider whether the possibility of costly signaling prior to game play will undermine the deterrence equilibrium in the baseline model. For the purposes of exposition all proofs are deferred to the end of the section.

Consider a variant of the model in which the challenger, when she transgresses, can also burn utility $c \geq 0$ (of her choosing) in order to signal information about her type. Let $\alpha\left(c\right)$ denote the

probability the defender responds to the transgression with war conditional on a costly signal of c.

We first state the main result – that when the military value of the transgression strictly exceeds its direct value, the previously-characterized deterrence equilibrium of our model is robust in the sense of satisfying universal divinity (Banks and Sobel 1987). There is thus a strong sense in which the ability to send costly signals does not undermine our main result.

Proposition C.8. *If* $\bar{\theta}_C^2 < \bar{\theta}_C^1 \iff \delta_C^m\left(\bar{\theta}_C^1\right) > \delta_C^d$, *then there exists a universally divine equilibrium in which*

- *the challenger never sends a costly signal, and transgresses i.f.f.* $\theta_C \geq \bar{\theta}_C^1$
- *the defender responds to the transgression with war* $(\alpha(c) = 1)$ *for all* $c < \max_{\theta_C \geq \bar{\theta}_C^1} \{\delta_C^m(\theta_C)\}$.

We now heuristically explain why when $\bar{\theta}_C^2 < \bar{\theta}_C^1$, there exists a universally divine equilibrium in which peaceful challengers cannot use costly signals to "separate" themselves and induce the defender to allow the transgression. Consider a strategy profile in which deterrence occurs and the challenger never sends the costly signal; the defender's inferences upon observing the transgression but no signal ($c = 0$) are therefore as in the original model, while transgressing and signaling ($c > 0$) is "off path." What must the defender believe about the challenger's intentions if he were to observe that she transgressed but also sent a costly signal $c > 0$? Roughly, universal divinity requires that he believe the signal to have been sent by the challenger type for whom successfully transgressing is most profitable, and respond accordingly. Crucially, this cannot be a peaceful type ($\theta_C < \bar{\theta}_C^2$). The reason is that when appeasement is impossible, there is some chance that the challenger is "opportunistically belligerent" $(\theta_C \in \left(\bar{\theta}_C^2, \bar{\theta}_C^1\right))$ – that is, initially deterrable, but would start a war if allowed to transgress. Since an opportunistically-belligerent challenger necessarily places a strictly greater value $w_C^2(\theta_C) - n_C^1$ on successfully transgressing than a peaceful one $n_C^2 - n_C^1 = \delta_C^d$, the defender must respond with war, and sending such a signal cannot be profitable for the challenger.

While the introduction of costly signaling does not cause the deterrence equilibrium to unravel when our main condition holds, the same cannot be said when the condition fails. The following proposition characterizes sufficient conditions for deterrence to collapse in all universally divine equilibria when the possibility for costly signaling is introduced.

Proposition C.9. *If* $\bar{\theta}_C^1 < \bar{\theta}_C^2 \iff \delta_C^d > \delta_C^m\left(\bar{\theta}_C^1\right)$ *and* $\delta_C^d > \max_{\theta_C \geq \bar{\theta}_C^2} \{\delta_C^m(\theta_C)\}$*, then in every universally divine equilibrium the challenger always transgresses.*

Thus, when appeasement is possible ($\bar{\theta}_C^1 < \bar{\theta}_C^2$), the potential for costly signaling can undermine the deterrence equilibrium. The reason is that the value $\delta_C^d = n_C^2 - n_C^1$ that peaceful challengers place on successfully transgressing is necessarily *higher* than the value $n_C^2 - w_C^1(\theta_C)$ that belligerent-but-appeasable challengers place on it. It is thus at least possible that a peaceful challenger places a higher value on transgressing than any unappeasably belligerent one ($\delta_C^d > \max_{\theta_C \geq \bar{\theta}_C^2} \{\delta_C^m(\theta_C)\}$) and can therefore "separate" themselves with costly signaling. An additional interesting implication of Propositions C.8 and C.9 is that when the transgression's military value is constant for all challenger types (as considered in Proposition 3), then with costly signaling there exists a universally divine equilibrium with deterrence if and *only if* $\delta_C^m \geq \delta_C^d$.

To conclude, we provide an example in which (i) the deterrence equilibrium exists in the baseline model, but (ii) with costly signaling deterrence unravels in every universally divine equilibrium.

Example: Suppose that $w_C^1(\theta_C) = \theta_C$, $\delta_C^m(\theta_C) = \delta_C^m \ \ \forall \theta_C$, $\theta_C \sim U[0, \theta_C^{\max}]$ where $\theta_C^{\max} > \bar{\theta}_C^1 = n_C^1$, and $\delta_C^d \in \left(\delta_C^m, \delta_C^m + \left(1-\bar{\beta}\right) \cdot \left(\bar{\theta}_C^{\max} - n_C^1\right)\right)$.

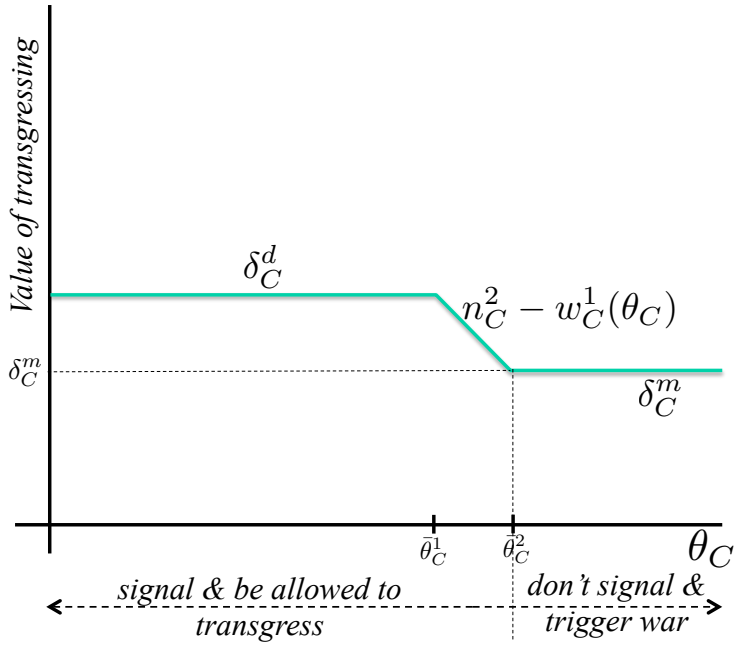
Here the deterrence equilibrium exists in the baseline model since $\beta = 1 - \frac{\delta_C^d - \delta_C^m}{\theta_C^{\max} - n_C^1} \geq \bar{\beta}$, but it unravels with costly signaling since $\delta_C^d > \delta_C^m$. The following strategy profile in which peaceful challengers separate themselves satisfies universal divinity.

- **(Challenger)** Always transgress, and $c(\theta_C) = \mathbf{1}_{\theta_C \leq \bar{\theta}_C^2} \cdot \delta_C^m$.

- **(Defender)** $\alpha(c) = \mathbf{1}_{c<\delta_C^m}$ (respond with war i.f.f. the costly signal is strictly less than δ_C^m).

The equilibrium is depicted in Figure 6. In it, peaceful and appeasable challengers $\theta_C \leq \bar{\theta}_C^2$ transgress and send a costly signal barely high enough ($c = \delta_C^m$) to distinguish themselves from unappeasably belligerent ones, and are therefore always allowed to transgress.[22] Unappeasably belligerent challengers $\theta_C > \bar{\theta}_C^2$ transgress without signaling, and always trigger a war.

Figure 6: Equilibrium with Costly Signaling



Proof of Proposition C.8 Let $\bar{\delta} = \max_{\theta_C \geq \bar{\theta}_C^1} \{\delta_C^m(\theta_C)\}$. Now consider a strategy profile-belief pair (σ, μ) (where $\mu(T|c)$ denotes the defender's interim-belief that θ_C is in the set T conditional on observing the transgression and costly signal c) that satisfies (i) the conditions of the proposition, (ii) $\mu\left(\arg\max_{\theta_C \geq \bar{\theta}_C^1} \{\delta_C^m(\theta_C)\} \middle| c\right) = 1 \ \forall c < \bar{\delta}$, and (iii) $\alpha(c)$ is a best response to $\mu(\cdot|c) \ \forall c \geq \bar{\delta}$. We argue that any pair (σ, μ) is a universally divine equilibrium.

We first introduce some additional useful notation. Let $\Pi(\theta_C, \alpha)$ denote the *net benefit* (excluding

[22] The figure uses $\delta_C^m = .2 < \delta_C^d = .3$, $\bar{\theta}^{\max} = 1$, and $n_C^1 = .5$.

signaling costs) of a challenger type θ_C deviating from her strategy in σ to transgressing and sending a signal that induces the defender to respond with war with probability α (so $\Pi(\theta_C, \alpha(c)) - c$ is the total net benefit of deviating to transgressing with a signal c). Then

$$\Pi(\theta_C, \alpha) = \begin{cases} (1-\alpha)\,\delta_C^d - \alpha\left(n_C^1 - w_C^1(\theta_C)\right) & \text{for } \theta_C \leq \bar{\theta}_C^2 \\ (1-\alpha)\left(w_C^2(\theta_C) - n_C^1\right) - \alpha\left(n_C^1 - w_C^1(\theta_C)\right) & \text{for } \theta_C \in \left[\bar{\theta}_C^2, \bar{\theta}_C^1\right] \\ (1-\alpha)\,\delta_C^m(\theta_C) & \text{for } \theta_C \geq \bar{\theta}_C^1 \end{cases}$$

It is easily verified that (i) $\Pi(\theta_C, \alpha)$ is *strictly* decreasing in α $\forall \theta_C$, and (ii) $\forall \alpha$, $\Pi(\theta_C, \alpha)$ is weakly increasing in θ_C over $\theta_C \leq \bar{\theta}_C^2$ and strictly increasing in θ_C over $\theta_C \in \left[\bar{\theta}_C^2, \bar{\theta}_C^1\right]$. This furthermore implies that $\Pi\left(\bar{\theta}_C^1, \alpha\right) > \Pi(\theta_C, \alpha)$ $\forall \theta_C < \bar{\theta}_C^2$.

The proof now proceeds in two steps. First, we argue that (σ, μ) is a Perfect Bayesian Equilibrium. Second, we argue that (σ, μ) survives the application of the universal divinity refinement, and is hence a universally divine equilibrium. To see that (σ, μ) is a PBE, observe that since $\bar{\theta}_C^2 < \bar{\theta}_C^1$, the deterrence equilibrium exists absent the potential for costly signaling, strategies and beliefs are as in the deterrence equilibrium on equilibrium path, and the defender's actions and beliefs satisfy sequential rationality by construction off-path. It remains only to show that no challenger type wishes to deviate to transgressing with a costly signal $c > 0$. Signals $0 < c < \bar{\delta}$ are clearly strictly worse than transgressing with no signal since they are costly but yield no increase in the probability the transgression will be allowed. Signals $c \geq \bar{\delta}$ are unprofitable since by previous observations about $\Pi(\theta_C, \alpha(c))$ we have $\Pi(\theta_C, \alpha(c)) \leq \max_{\theta_C \geq \bar{\theta}_C^1} \{\Pi(\theta_C, \alpha(c))\} = (1 - \alpha(c)) \cdot \max_{\theta_C \geq \bar{\theta}_C^1} \{\delta_C^m(\theta_C)\} \leq \bar{\delta} \leq c$.

Finally, to see (σ, μ) survives universal divinity, we argue that it survives the iterative application of the NWBR signaling criterion, which is a strengthening of D2 (Fudenberg and Tirole 1991, pp. 454). First observe that all $\alpha \in [0, 1]$ are in the set of defender mixed best responses to the original type space Θ. Second, observe that for $c \geq \bar{\delta}$, $\Pi(\theta_C, \alpha) - c \leq 0$ $\forall (\theta_C, \alpha)$; since no type can be made strictly better off sending such signals for *any* value of α, no type may be eliminated through the

application of NWBR; the associated beliefs in (σ, μ) therefore satisfy universal divinity.

Last consider $c \in \left(0, \bar{\delta}\right)$. For any pair $\hat{\theta}_C \in \arg\max_{\theta_C \geq \bar{\theta}_C^1} \{\delta_C^m(\theta_C)\}$ and $\theta_C \notin \arg\max_{\theta_C \geq \bar{\theta}_C^1} \{\delta_C^m(\theta_C)\}$ we have that $\Pi\left(\hat{\theta}_C, 0\right) - c = \bar{\delta} - c > 0$, so there exists a mixed best response ($\alpha = 0$) that would make type $\hat{\theta}_C$ strictly benefit from the deviation). In addition, $\Pi\left(\hat{\theta}_C, \alpha\right) > \Pi(\theta_C, \alpha)\ \forall \alpha > 0$. Hence any α that would make θ_C indifferent to sending c (which must be > 0) would make $\hat{\theta}_C$ strictly prefer to send c, and θ_C may be pruned. Therefore beliefs $\mu\left(\arg\max_{\theta_C \geq \bar{\theta}_C^1} \{\delta_C^m(\theta_C)\} \middle| c\right) = 1$ result from the first application of NWBR, and moreover yield a *unique* mixed best response $\alpha = 1$ that is exactly the defender's strategy in our profile. Since no type profits from deviating in this profile, further applications of NWBR cannot further restrict beliefs, and the profile satisfies universal divinity. ■

Proof of Proposition C.9 Let $\bar{\delta} = \max_{\theta_C \geq \bar{\theta}_C^2} \{\delta_C^m(\theta_C)\} < \delta_C^d$. Now suppose the conditions hold and consider a universally divine equilibrium. Transgressing and not signaling ($c = 0$) strictly dominates not transgressing for all $\theta_C > \bar{\theta}_C^1$, so all such types must transgress. Next consider types $\theta_C \leq \bar{\theta}_C^1 < \bar{\theta}_C^2$. We argue for any $\hat{c} \in \left(\bar{\delta}, \delta_C^d\right)$, universal divinity implies $\alpha(\hat{c}) = 0$ (the defender always allows the transgression), implying that transgressing and sending $\hat{c}$ yields net benefit of $\delta_C^d - \hat{c} > 0$ over not transgressing for such types, implying that they also must transgress in equilibrium.

Observe that for types $\theta_C \geq \bar{\theta}_C^2$, transgressing and sending the costly signal $\hat{c}$ is strictly dominated by transgressing and sending no signal; the *best* payoff the former could yield is $w_C^2(\theta_C) - \hat{c}$ (if it results in the transgression always being allowed) while the *worst* payoff the latter could yield is $w_C^1(\theta_C)$ (if it results in the transgression never being allowed), and $w_C^1(\theta_C) > w_C^2(\theta_C) - \hat{c} \iff \hat{c} > \delta_C^m(\theta_C)$. Thus, if $\hat{c}$ is on equilibrium path then only challengers types $\theta_C < \bar{\theta}_C^2$ may be sending $\hat{c}$. If $\hat{c}$ is off equilibrium path then (the first iteration of) universal divinity eliminates types $\theta_C \geq \bar{\theta}_C^2$ from the defender's beliefs since they would benefit from the deviation for no defender responses, while types $\theta_C < \bar{\theta}_C^1$ would benefit for some defender responses. In either case this is sufficient to imply that the defender's best response must be to allow the transgression. ■